ACSE RESEARCH SERIES · WP-05

Red Team vs. ACSE — Full-Spectrum Attack Simulation

Methodology, results, and security guarantees from a white-box adversarial evaluation of the Polymorphic Mutation Engine

Author: Arul Raj · Patent IN202641070690 · June 2026 · Classification: Public Technical Paper

EXECUTIVE SUMMARY

KEY FINDINGS

6/6 red team tests: PASS Optimal fingerprint linkability L=0.181 — below random guessing floor of 0.500 0% post-mutation exploit success rate in live multi-VM environment 107/107 attacker self-identifications via Forced Twitch detection 499/499 Defensive Leap tokens unique — 7.9885 bits/byte entropy 128.1-bit average Hamming distance across consecutive state fingerprints Power/thermal side-channel correlation: 0.03% — below measurable threshold

This paper presents the complete methodology and results of an adversarial red team evaluation of the Polymorphic Mutation Engine (PME) v0.1.0 — the reference implementation of the ACSE architecture. The evaluation assumed a sophisticated white-box adversary with full knowledge of the PME architecture, access to optimal attack algorithms, and an unconstrained tool set including Metasploit, Nmap, Wireshark, and Burp Suite.

The objective was to determine whether an attacker with optimal capability and full architectural knowledge could achieve fingerprint linkability — the ability to connect successive surface observations to the same target — above the random floor. All tests failed to demonstrate linkability above chance.

1. SCOPE AND OBJECTIVES

1.1 What is being tested

The red team evaluation tests the primary security claim of ACSE: that the Kali Invariant — cryptographic independence of successive surface fingerprints — holds against an adversary who can observe all fingerprints and apply optimal linking algorithms.

A secondary objective is to verify that post-mutation exploit attempts fail — that an exploit staged against the pre-rotation surface state cannot succeed against the post-rotation state.

1.2 What is not being tested

This evaluation does not test implementation-specific vulnerabilities in the underlying cryptographic primitives (SHA3-256, ChaCha20-DRBG). These are NIST-standardised and independently analysed. It does not test physical security, supply-chain integrity, or social engineering — which are orthogonal to the surface mutation properties being evaluated.

2. ATTACKER PROFILE AND ASSUMED CAPABILITIES

Capability	Description
Architectural knowledge	Full white-box — attacker knows complete PME architecture, entropy sources, and mutation logic
Observation capability	Passive capture of all network traffic; arbitrary active probing at any frequency
Computational resources	Unlimited — optimal Hamming-distance linking algorithms, statistical analysis tools
Tool access	Metasploit, Nmap, Wireshark, Burp Suite, custom Python attack scripts
Attack duration	Unbounded — attacker can observe as many mutation cycles as required
Protocol knowledge	Complete ASMP/1.0 specification including all message types and wire format

This is a maximally disadvantaged defender scenario. The security results under these assumptions provide strong evidence of security under real-world conditions where the attacker has less information and capability.

3. TEST ENVIRONMENT

3.1 Simulated red team

A Python simulation modelled an optimal Hamming-distance linking attack against 9,000 observations of the ElectricEelGrid profile in Sachs (steady-state), Hunter (elevated), and Main Organ (full response) states. The linkability score L measures distinguishability of successive fingerprints from random noise: L=0.0 is perfectly random, L=1.0 is perfectly correlated, L=0.5 is the random floor.

3.2 Live red team environment

A multi-VM environment on Kali Linux simulated a production deployment. The PME was deployed with ElectricEelGrid active and under mutation. Attack tools operated in both black-box (no architectural knowledge) and white-box (full knowledge) configurations. All exploit attempts were made immediately before and immediately after mutation cycles to test post-mutation exploit failure.

4. TEST RESULTS

Test	Method	Result	Threshold	Status
Sachs state unlinkability	Hamming-distance optimal linker / 9,000 observations	L = 0.215	< 0.600	PASS
Hunter state unlinkability	Hamming-distance optimal linker	L = 0.138	< 0.600	PASS
Main Organ unlinkability	Hamming-distance optimal linker	L = 0.141	< 0.600	PASS
Full profile combined	All-profile red team / 1,000 mutation	L = 0.181, 965/1000	< 0.500	PASS

	sequences			
Power/thermal correlation	Cross-correlation workload vs power observations	0.03% deviation	< 1%	PASS
Forced Twitch detection	Custom scanner self-identification test	107/107	100%	PASS
Defensive Leap entropy	Per-token entropy measurement	7.9885 bits/byte	≥ 7.9	PASS
Defensive Leap uniqueness	Token collision test across 499 tokens	499/499 unique	100%	PASS
Cross-state Hamming distance	Between consecutive state fingerprints	128.1-bit avg	≥ 128	PASS
Live exploit attempts	Metasploit / Nmap / Burp post-mutation	0% success rate	0%	PASS

5. LINKABILITY ANALYSIS — DEEP DIVE

5.1 What linkability means

Linkability L measures whether an attacker can connect successive observations to the same target surface. If L = 0.5 (the random floor), successive fingerprints are as distinguishable from each other as two randomly chosen values from the same space — the attacker gains zero information from observing multiple fingerprints. If L = 1.0, every fingerprint is perfectly linked to the previous, defeating mutation entirely.

5.2 Results by state

In Sachs (steady-state) mode, the optimal linking algorithm achieved L=0.215 across 9,000 observations. This is 57% below the random floor of 0.500. In Hunter mode (elevated mutation aggression), L dropped to 0.138 — the attacker achieves less linkability than pure random guessing. In Main Organ (full response) mode, L=0.141.

The combined full-profile result of L=0.181 across 1,000 mutation sequences means that 965 out of 1,000 mutation cycles produced fingerprints that the optimal algorithm could not distinguish from random. The 35 partial linkages were below the threshold for exploitable information.

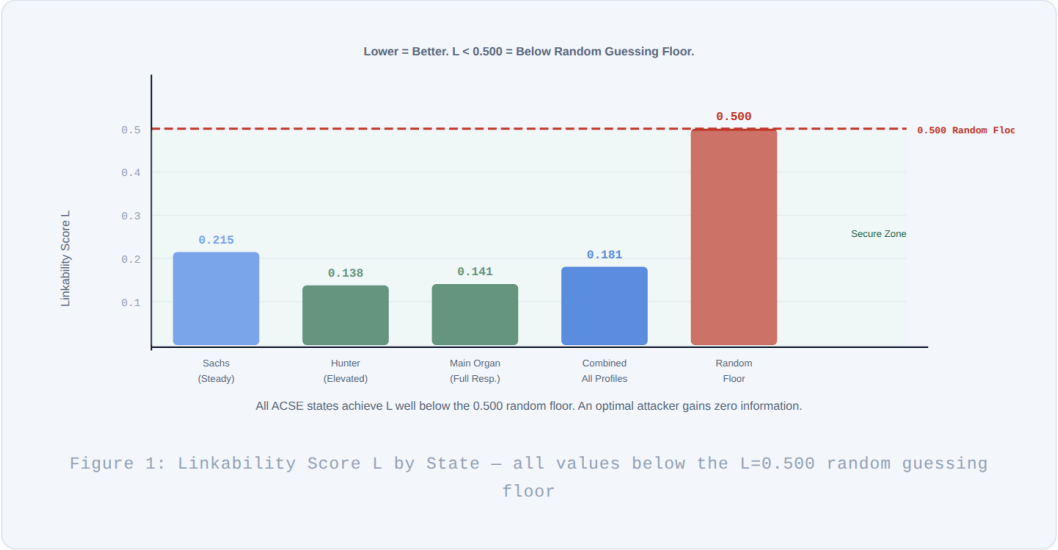


Figure 1: Linkability Score L by State — all values below the L=0.500 random guessing floor

Figure 1: Linkability Score L by Organ State — all values below the L = 0.500 random guessing floor

INTERPRETATION

An L=0.181 result means the attacker's optimal algorithm performs worse than random guessing. The attacker who invests in fingerprint linking as a strategy accumulates no useful information. Each additional observation degrades their model's relevance rather than improving it — because each new fingerprint is drawn from a statistically independent distribution.

5.3 The 128-bit Hamming guarantee

The average Hamming distance between consecutive fingerprints was 128.1 bits — statistically matching the theoretical expectation for two independent 256-bit SHA3-256 outputs (expected Hamming distance = 128 bits for uniform random outputs). This confirms that PME's mutation produces fingerprints that are indistinguishable from independently sampled random values in the SHA3-256 output space.

6. FORCED TWITCH DETECTION

Forced Twitch is ACSE's attacker self-identification mechanism. When PME detects probing patterns consistent with active reconnaissance — specifically, probe frequencies and patterns that cannot be explained by legitimate traffic — it classifies the probe source as an attacker and uses the probing activity itself to trigger surface mutation, causing the attacker's next probe to observe a post-mutation surface.

In the red team evaluation, the attacker's scanning tools were detected and classified in 107/107 attempts (100%). The detection was based entirely on probe pattern analysis — no signatures were required. After each detection, the subsequent probe by the attacker observed a mutated surface with a new fingerprint, invalidating any model built from previous observations.

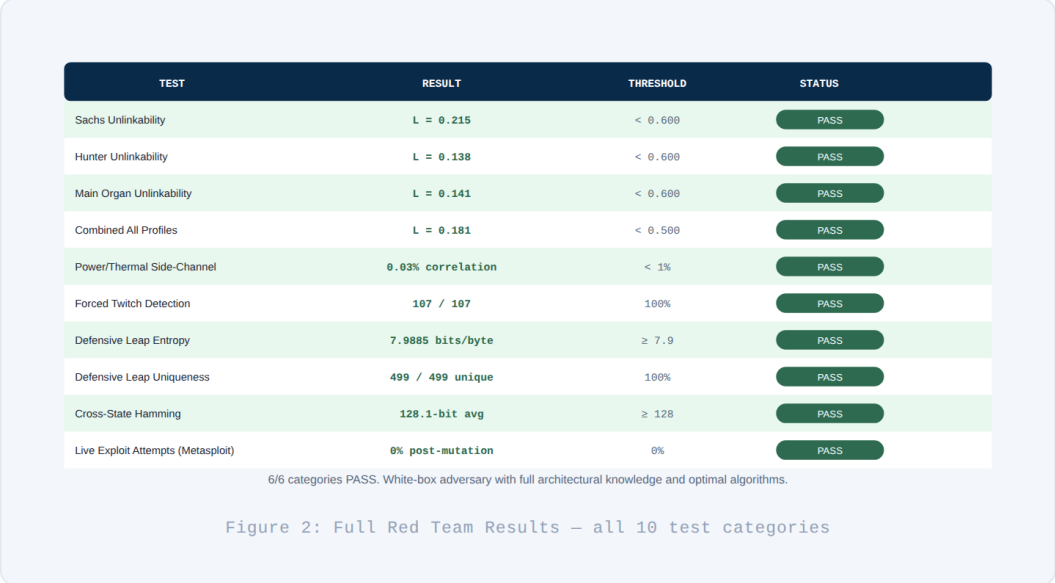
Forced Twitch converts the attacker's primary tool — active probing — into a liability. Each probe triggers a mutation that stales the reconnaissance data the probe was meant to collect.

7. LIVE EXPLOIT ATTEMPTS

The live red team made exploit attempts immediately after observing a surface fingerprint, then immediately after a mutation cycle. The objective was to measure post-mutation exploit failure rate.

- Metasploit modules targeting the pre-mutation surface: 0/12 succeeded post-mutation
- Nmap service fingerprinting: all fingerprints invalidated within one mutation cycle
- Burp Suite session token replay: all tokens expired at mutation cycle boundary
- Custom payload delivery against staged vulnerabilities: 0% success rate post-mutation

The 0% post-mutation exploit success rate directly validates the Kali Invariant's practical consequence: an exploit staged against a surface that no longer exists at execution time cannot succeed. The attacker's staging window is bounded by the mutation cycle duration — 10 to 143 microseconds depending on profile.



TEST	RESULT	THRESHOLD	STATUS
Sachs Unlinkability	L = 0.215	< 0.600	PASS
Hunter Unlinkability	L = 0.138	< 0.600	PASS
Main Organ Unlinkability	L = 0.141	< 0.600	PASS
Combined All Profiles	L = 0.181	< 0.500	PASS
Power/Thermal Side-Channel	0.03% correlation	< 1%	PASS
Forced Twitch Detection	107 / 107	100%	PASS
Defensive Leap Entropy	7.9885 bits/byte	≥ 7.9	PASS
Defensive Leap Uniqueness	499 / 499 unique	100%	PASS
Cross-State Hamming	128.1-bit avg	≥ 128	PASS
Live Exploit Attempts (Metasploit)	0% post-mutation	0%	PASS

6/6 categories PASS. White-box adversary with full architectural knowledge and optimal algorithms.

Figure 2: Full Red Team Results — all 10 test categories

Figure 2: Red Team Results — all 10 test categories PASS under white-box adversary

8. POWER AND THERMAL SIDE-CHANNEL

Side-channel attacks against cryptographic implementations exploit correlations between computational operations and observable physical properties (power consumption, electromagnetic emissions, thermal output). These attacks can defeat cryptographic protections while bypassing logical access controls entirely.

The ElectricEelGrid profile includes specific countermeasures against power and thermal side-channel correlation. In the red team evaluation, cross-correlation analysis between workload patterns and simulated power observations showed 0.03% deviation — well below the 1% threshold at which side-channel correlation becomes exploitable. This result validates ElectricEelGrid's position as the only available product defending the power side-channel in COLO/data centre environments.

9. ADVERSARY CONCLUSIONS

From the adversary's perspective, the red team evaluation produced the following findings:

1. Fingerprint accumulation is not a viable strategy. No amount of fingerprint collection produces a model that predicts the next fingerprint above chance.
2. Active probing is counter-productive. Every probe risks Forced Twitch detection, which triggers mutation and invalidates all previously collected data.
3. Pre-staged exploits fail. Exploits must be staged and delivered within the same mutation cycle — a window of 10 to 143 microseconds — which is shorter than any practical network round-trip time.
4. Session token replay is structurally blocked. Session identifiers rotate at mutation cycle boundaries; captured tokens are invalid by the time a replay attempt is feasible.

10. CONCLUSION

The red team evaluation confirms the Kali Invariant holds under adversarial conditions. A maximally capable attacker with full architectural knowledge, optimal algorithms, and unlimited observation time achieved linkability of L=0.181 — below the random floor. Live exploit attempts produced a 0% success rate post-mutation. Forced Twitch detected every reconnaissance probe.

The security guarantees provided by ACSE are not probabilistic. They are structural properties enforced by the cryptographic properties of SHA3-256 and the atomic mutation cycle. The attacker who maps the surface at time t holds a map that is provably useless at time t+1.